Discussion Paper No. 892R

Refining Cheap-Talk Equilibria*

by
STEVEN A. MATTHEWS
Northwestern University

MASAHIRO OKUNO-FUJIWARA
University of Tokyo

ANDREW POSTLEWAITE
University of Pennsylvania

June 1990
revised November 1990

This paper is both a Discussion Paper at the Center for Mathematical Studies in Economics and Management Science, Northwestern University, and a Working Paper at the Center for Analytic Research in Economics and the Social Sciences, University of Pennsylvania.

---

* This paper is taken from the first part of "Communication in Bayesian Games: Issues and Problems," by the same authors, which was presented at the California Institute of Technology, December, 1988, and at the NBER-NSF Decentralization Conference, April, 1989. We thank, for their comments, an anonymous referee, Michael Chwe, Preston McAfee, Tom Palfrey, Matthew Rabin, Joel Sobel, Yossi Spiegel, and Asher Wolinsky. The first and third authors acknowledge support from the National Science Foundation.

Title: Refining Cheap-Talk Equilibria

Authors: Steven A. Matthews, Masahiro Okuno-Fujiwara,
and Andrew Postlewaite

Correspondent: Steven A. Matthews
Department of Economics
Northwestern University
Evanston, IL 60208

Running Head: Refining Cheap-Talk Equilibria

Abstract:

Several conceptual points are made concerning communication in games of asymmetric information. Equilibrium refinements of Sender-Receiver cheap-talk games that are based on the concept of a putative equilibrium, and which rely on the presence of a rich language with literal meanings, are discussed. Three nested criteria are proposed: strong announcement-proofness, announcement-proofness, and weak announcement-proofness.

JEL Classification Numbers: 026

Keywords: cheap talk, equilibrium refinements

1. Introduction

Studies of cheap-talk communication must face the refinement issue. Allowing cheap talk, i.e., allowing players to exchange costless, nonverifiable messages, necessarily expands a game's set of equilibrium outcomes: every original outcome is achieved by a cheap-talk equilibrium in which the players babble uninformatively and listen inattentively. In some games these uncommunicative equilibria appear plausible, such as when the players have opposing interests. However, it is generally felt that in games in which interests are more aligned, uncommunicative equilibria are less likely to be played than are equilibria involving meaningful, "coordinating" communication. Our goal in this paper is to discuss and to formalize this intuition in terms of an equilibrium refinement criterion, which should serve to clarify this intuition, establish its validity, and guide its systematic application.

Standard refinement criteria have little force in cheap talk games. Every equilibrium outcome of a cheap talk game is achieved by an equilibrium in which every message is sent with positive probability (assuming a countable message space). Every information set in the communication part of the game can be "on-the-equilibrium-path." Consequently, in the Sender-Receiver games we consider, the set of equilibrium outcomes is unaffected by standard refinement criteria.[1]

Our interest is in games of asymmetric information, where communication can serve to convey a player's private information. The simplest of these games is a Sender-Receiver cheap-talk game, as introduced by Crawford and Sobel [6]. Only the Sender has private information in such a game, and only the Receiver has payoff-relevant actions. These games are not without application,[2] and they provide a good testing ground for ideas about cheap talk and refinements.

---

[1] Such as those of Banks and Sobel [2], Cho and Kreps [5], Grossman and Perry [11], Kohlberg and Mertens [12], McLennan [15], or Okuno-Fujiwara and Postlewaite [17].

[2] For example, Stein [21]. Matthews [14] and Forges [10] use variants of Sender-Receiver games.

Farrell [7, 8] provided the first refinement criterion for Sender-Receiver cheap-talk games. His contribution was fundamental, providing the central concept of a rich language with literal meanings through which the Sender can upset a putative equilibrium. His "neologism-proof" criterion is discussed in detail in in Section 3.

Two criticisms of this criterion are presented in Section 4. One concerns its inability to reject uncommunicative equilibria in some coordination games, and the other concerns an inconsistency in its rationale.

Three new criteria are presented and discussed in Section 5: the strongly announcement-proof criterion, the announcement-proof criterion, and the weakly announcement-proof criterion. Their rationales extend that of neologism proofness.

In Section 6 we indicate that the two new criteria reject uncommunicative equilibria, and accept communicative equilibria, in certain kinds of coordination games.

Before starting, we should mention another refinement criterion making use of a language with literal meanings, the "coherent plans" criterion of Myerson [16]. It differs significantly from ours and Farrell's in two ways: it relies on a mediator to filter communications, and it judges the plausibility of statements relative to a reference allocation that need not be an equilibrium outcome. Our focus on face-to-face, unmediated communication has led us down a different path.[3]

There are no "big theorems" in this paper. It should be viewed as a discussion of some of the issues involved in modeling communication, together with illustrative examples and definitions. For easy reference, the examples are summarized in the Appendix.

2. Sender-Receiver Games

A general, formal way of looking at cheap talk communication is to consider adding k stages of pre-play message exchange to a game Γ^0 to form a new game Γ^k. In order to ease subsequent generalization, it is useful to view a Sender-Receiver game in this formal way,

---

[3] We recently became aware of independent and concurrent work, by Rabin [19], in which a more closely related criterion is proposed. See the Postscript.

even though the base game Γ^0 is a trivial one-player game. We write the base game as a function of the probability distribution on the Sender's type, since the effect of communication will be to change the Receiver's perception of this distribution. Thus, the base game is $\Gamma^0(\pi) = \langle T, \pi, A, u_S, u_R \rangle$, where

T = finite set of Sender types;

A = finite set of actions available to the Receiver;

π = the Receiver's beliefs about the Sender's type, $\pi \in \Delta T$;[4]

u_S = the Sender's payoff function, $u_S: A \times T \to \Re$;

u_R = the Receiver's payoff function, $u_R: A \times T \to \Re$.

A mixed strategy played by the Receiver, $\sigma \in \Delta A$, results in the interim expected payoffs,

$$(2.1) \qquad u_S(\sigma \mid t) = \sum_{a \in A} \sigma(a) u_S(a,t),$$

$$(2.2) \qquad u_R(\sigma \mid \pi) = \sum_{t \in T} \sum_{a \in A} \pi(t)\sigma(a) u_R(a,t).$$

An equilibrium of $\Gamma^0(\pi)$ is a strategy σ that maximizes the Receiver's expected payoff, $u_R(\sigma \mid \pi)$, over ΔA. The set of equilibria of $\Gamma^0(\pi)$ is denoted $E^0(\pi)$. It is useful to view E^0 as a correspondence on ΔT, since the effect of cheap talk will be to change the Receiver's beliefs about the Sender's type.

The Sender-Receiver game based on $\Gamma^0(\pi)$ is obtained by appending a stage in which the Sender sends a message to the Receiver. The set of available messages is M, assumed to be countably infinite. A (behavioral) strategy for the Sender, referred to as a *talking strategy*, is a function $\tau: T \to \Delta M$. A strategy for the Receiver, referred to as an *action strategy*, is a function $\alpha: M \to \Delta A$. The strategy sets are S_S and S_R, respectively. The payoff of a type t Sender, given a strategy pair (τ,α), is

$$(2.3) \qquad u_S(\tau,\alpha \mid t) = \sum_{m \in M} \sum_{a \in A} \tau(m \mid t)\alpha(a \mid m) u_S(a,t).$$

The corresponding payoff for the Receiver is

---

[4] For any set X, ΔX is the set of probability distributions on X.

$$(2.4) \qquad u_R(\tau,\alpha) = \sum_{t\in T}\sum_{m\in M}\sum_{a\in A}\pi(t)\tau(m\mid t)\alpha(a\mid m)u_R(a,t).$$

The Sender-Receiver game just defined is $\Gamma^1(\pi) = \langle T, \pi, S_S, S_R, u_S, u_R\rangle$. A strategy pair (τ,α) is a *Bayesian equilibrium* of $\Gamma^1(\pi)$ if

$$(2.5) \qquad u_R(\tau,\alpha) \geq u_R(\tau,\hat{\alpha}) \quad \text{for all } \hat{\alpha}\in S_R, \text{ and}$$

$$(2.6) \qquad u_S(\tau,\alpha\mid t) \geq u_S(\hat{\tau},\alpha\mid t) \quad \text{for all } t\in T \text{ and } \hat{\tau}\in S_S.$$

It is usual to restrict attention to equilibria satisfying a perfection property. Notice that if α is an action strategy and m is a message, then $\alpha(m)\in\Delta A$ is a strategy for the Receiver in the underlying game, $\Gamma^0(\pi)$. Sequential rationality requires $\alpha(m)$ to be an equilibrium of $\Gamma^0(\hat{\pi})$, where $\hat{\pi}$ denotes the Receiver's beliefs about the Sender's type after receiving message m. Thus, we introduce a *belief function*, $\beta: M\to\Delta T$, by which the Receiver updates his beliefs. A triple (τ,α,β) is *sequentially rational* if

$$(2.7) \qquad \alpha(m)\in E^0(\beta(m)) \quad \text{for all } m\in M.$$

The triple is *Bayes-consistent* if β satisfies Bayes' rule whenever possible:

$$(2.8) \qquad \beta(t\mid m) = \frac{\pi(t)\tau(m\mid t)}{\sum_{s\in T}\pi(s)\tau(m\mid s)} \quad \text{for all } m\in\tau(T).^5$$

The triple (τ,α,β) is a *perfect Bayesian equilibrium* if (τ,α) is a Bayesian equilibrium and (τ,α,β) is sequentially rational and Bayes-consistent; equivalently, (τ,α,β) is a perfect Bayesian equilibrium if it satisfies (2.6)–(2.8). Henceforth, "equilibrium" refers to perfect Bayesian equilibria. The set of equilibria of $\Gamma^1(\pi)$ is denoted $E^1(\pi)$.

We define an "outcome" only in terms of what is payoff-relevant, i.e., in terms of actions and types, not messages. Thus, an *outcome* of $\Gamma^0(\pi)$ or $\Gamma^1(\pi)$ is a mapping $o: T\to\Delta A$. The sets of equilibrium outcomes are, respectively,

---

[5] For $S\subset T$, we define $\tau(S) = \{m\in M \mid \tau(m\mid t)>0 \text{ for some } t\in S\}$. In words, $\tau(S)$ is the set of messages sent by the types of Sender in S.

$$(2.9)\qquad O^0(\pi) = \{o: T\to\Delta A \mid \text{for some } \sigma\in E^0(\pi),\ o(a\mid t)=\sigma(a) \text{ for all } a\in A \text{ and } t\in T\},$$

$$(2.10)\qquad O^1(\pi) = \{o: T\to\Delta A \mid \text{for some } (\tau,\alpha,\beta)\in E^1(\pi),\ o(a\mid t)=\sum_{m\in M}\alpha(a\mid m)\tau(m\mid t) \text{ for all } a\in A \text{ and } t\in T\}.$$

As is well known, adding communication does not destroy equilibrium outcomes:

$$(2.11)\qquad O^0(\pi)\subseteq O^1(\pi).^{6}$$

The proof of (2.11) is based on the observation that an uninformative talking strategy and a belief function that treats all messages as though they were uninformative are best responses to each other. Thus, given $\sigma\in E^0(\pi)$, a triple (τ,α,β) is an equilibrium of $\Gamma^1(\pi)$ achieving the same outcome as σ if $\tau(\cdot\mid t)=\tau(\cdot\mid s)$ for all $s,t\in T$, $\alpha(m)=\sigma$ for all $m\in M$, and $\beta(m)=\pi$ for all $m\in M$. Such equilibria in which τ is a constant function have been called *babbling equilibria*.

In non-babbling equilibria, the Sender communicates informatively. Nonetheless, his communication is ineffective if it fails to induce the Receiver to take actions different than he would have taken in the absence of communication, even though it alters his beliefs. It seems inappropriate to refer to an equilibrium with ineffective communication as a "communication equilibrium." Thus, say that an equilibrium (τ,α,β) of $\Gamma^1(\pi)$ is a *communication equilibrium* only if it gives rise to an outcome $o\in O^1(\pi)\setminus O^0(\pi)$, i.e., an outcome that is not an equilibrium outcome in the absence of a message stage. Equilibria for which this is not true are *no-communication equilibria*.[7]

---

[6] It can be shown that $O^0(\pi)=\{o\in O^1(\pi) \mid o(s)=o(t) \text{ for all } s,t\in T\}$.

[7] Although not our topic in this paper, we note that *communication impervious* games, i.e., games in which adding communication does not create new equilibrium outcomes, are quite interesting. A descriptive analysis of such a game need not be altered if it is learned that pre-play communication is possible, except in so far as it may allow the players to coordinate on a particular equilibrium. On the other hand, the problem of designing an incentive scheme should include communication imperviousness as a constraint if the players cannot be prohibited from pre-play communication. Some results along these lines are available: Crawford and Sobel [6] and Seidmann [20] find sufficient conditions for a Sender-Receiver game to be communication-impervious, and Palfrey and Srivistava [18]

On the other hand, the interesting Sender-Receiver games have communication equilibria. In such games, both the Sender and the Receiver want some of the former's information transmitted to the latter. For a given type, the Sender and Receiver have similar preferences over actions, but their preferences change significantly with the type. In these situations the Sender's information can coordinate the Receiver's action.

3. Neologism-Proof Equilibria

Intuition often suggests a selection between communication and no-communication equilibria. However, as Farrell [7, 8] observes, standard refinement criteria have no force in cheap talk games. Accordingly, he devised the "neologism-proof" criterion for cheap-talk Sender-Receiver games.[8] We now summarize this criterion and its rationale.

A motivating example is the game Γ_1 (adapted from Farrell, [8]).

	action A	action B	action C
type a	2, 3	0, 0	1, 2
type b	0, 0	2, 3	1, 2

Game Γ_1

As in each of our examples, the Sender's types are equally likely. In Γ_1 he has two types, a and b, and the Receiver has three actions, A, B, and C. The first of the two payoffs in each cell is the Sender's, the second is the Receiver's.

The no-communication equilibrium outcome of Γ_1 is $o(\cdot) = C$.[9] There is also a communication outcome, $o(a) = A$ and $o(b) = B$. This outcome can be achieved by using two

---

construct incentive efficient mechanisms for a particular environment which are communication-impervious.

[8] The neologism-proof criterion can be extended to cheap talk games that are not Sender-Receiver games; see, e.g., Austen-Smith [1], Farrell and Gibbons [9], or Matthews [13]. Grossman and Perry [11] propose a refinement similar in spirit to Farrell's for general extensive form games. The criteria suggested in this paper could be extended to more general games, most obviously to signaling games.

[9] We abuse notation for deterministic outcomes. E.g., $o(t) = C$ instead of $o(C \mid t) = 1$.

messages, with the type *a* Sender sending a message meaning, in equilibrium, 'I am type *a*,' and type *b* sending a message meaning, 'I am type *b*.' The Sender's information plays a purely coordinating role, since the Sender and the Receiver have the same preferences over actions. The communication outcome certainly appears the more plausible of the outcomes.

Why does the no-communication outcome of Γ_1 seem implausible? Consider, for example, the *random babbling equilibrium* in which both types of Sender select a message according to a uniform distribution over *M*, and the Receiver chooses *C* in response to any message. Common sense suggests that the Sender should communicate with the Receiver in order to achieve an outcome better for them both. Rather than babble, the Sender should make a speech along the following lines when, for example, his type is *a*:

> "I am type *a*. You should believe me, since if you do you will take action *A*, which is an action that I would prefer over the action which you have been planning to take, *C*, if and only if my true type is *a*."

This speech is compelling, as is the corresponding one for type *b*. The Receiver should reasonably be expected to believe these speeches, and so take action *A* (*B*) if the Sender is type *a* (*b*). In this case the speeches prevent the babbling equilibrium from being played.

The displayed speech has a literal meaning, which includes, "My type is *a*." In traditional game theory, a message has no literal meaning. Messages are given meaning only by an equilibrium; a message's equilibrium meaning is determined by the equilibrium beliefs it induces. In the random babbling equilibrium, the equilibrium meaning of the displayed speech is, "this speech is the result of random babbling and you should make no inferences from it," which is a meaning at odds with its literal meaning.

Farrell's [7, 8] first contribution is to recognize that many messages have literal meanings; the neologism-proof criterion takes the motivating "speeches" very seriously. Of course, the literal meaning of a message may actually be its equilibrium meaning as determined in an all-encompassing "game of life." Alternatively, the literal meaning of a message may be determined by the meanings of its individual components ("words"), and the

application of a commonly known grammar to their combination. The literal meaning of a message constitutes a focal belief for it to induce. Farrell assumes that for every nonempty subset K of types, messages exist that have the literal meaning, "my type is in K."

Farrell then develops a theory of credibility to determine when the literal meaning of a message should be believed. It starts with a putative equilibrium to be tested; literal meanings are declared credible relative to it. A message with the literal meaning, "my type is in K," is a *neologism* relative to the putative equilibrium if it is sent with zero probability in that equilibrium. Being a neologism is one of two conditions that must be satisfied for a message with a literal meaning to be credible. (If the message is sent with positive probability in the equilibrium, then either its literal meaning is not credible because it conflicts with the meaning it is imbued with by Bayes' rule, or the two meanings are the same so that the literal meaning cannot upset the equilibrium.)

Formally, the requirement that a credible message be a neologism is severe: every equilibrium outcome in any cheap talk game can be supported by an equilibrium in which every message is sent with positive probability (if M is countable). Farrell makes an extra-theoretical assumption at this point. He simply assumes that in every relevant equilibrium, for every nonempty subset K of types, a message with the literal meaning "my type is in K" is not sent, i.e., is a neologism. Although this assumption does not square with traditional game theory, it is plausible, at least when the set of messages M is infinite — it can perhaps be put on a firm basis if players can adopt only strategies of finite complexity, thereby ruling out mixed strategies with infinite supports.

The second condition for a message with the literal meaning "my type is in K" to be believed is for precisely the Sender types in K to prefer (strictly) this message to be believed over what they would get in the putative equilibrium. That is, if the statement were believed in the sense that the Receiver responded to it by taking an action that is optimal for him when his beliefs are conditioned on the event $\{t \in K\}$, then every type in K would strictly prefer to make the statement than to play the equilibrium, and every type

not in K would weakly prefer to stay with the equilibrium. If a neologism satisfies this second condition, it is a *credible neologism* relative to the putative equilibrium.

Farrell's theory of credibility is that the literal meaning of a credible neologism should override its equilibrium meaning. But the equilibrium is destroyed if it is altered by making the Receiver believe a credible neologism, since the types in K would then prefer to send the neologism instead of their equilibrium messages. Thus, the putative equilibrium fails the test if a credible neologism exists relative to it. A *neologism-proof equilibrium* is an equilibrium relative to which a credible neologism does not exist.

In game Γ_1, the neologism, "my type is a," is credible relative to the no-communication equilibrium. This statement results in A if the Receiver believes it, and only type a prefers A to the no-communication outcome, C. Thus, the no-communication equilibrium is not neologism proof. The communication equilibrium is neologism proof; no neologism is credible relative to it because it gives both types of Sender their maximum payoffs.

Formally, although its rationale depends on these concepts, the neologism-proof criterion can be defined without reference to literal meanings and unsent messages.

DEFINITION 3.1: For each nonempty $K \subseteq T$, a *neologism* is an object represented as 'K.' The neologism is *believed* if it causes the Receiver to adopt the beliefs $\hat{\beta}('K')$ defined by

$$(3.1) \qquad \hat{\beta}(t \mid 'K') = \begin{cases} \dfrac{\pi(t)}{\sum_{s \in K} \pi(s)} & \text{if } t \in K \\ 0 & \text{if } t \notin K\,. \end{cases}$$

A neologism 'K' is *credible relative to an equilibrium* (τ,α,β) of $\Gamma^1(\pi)$ if two conditions hold:

C1′: $u_S(\sigma \mid t) > u_S(\tau,\alpha \mid t)$ for all $t \in K$ and $\sigma \in E^0(\hat{\beta}('K'))$,

C2′: $u_S(\sigma \mid t) \leq u_S(\tau,\alpha \mid t)$ for all $t \notin K$ and $\sigma \in E^0(\hat{\beta}('K'))$.[10]

---

[10] This is not quite Farrell's [7, 8] definition, as he assumes that $E^0(\hat{\beta}('K'))$ contains a unique equilibrium for each nonempty set K of types.

The equilibrium (τ,α,β) is *neologism proof* if no neologism is credible relative to it. An equilibrium outcome o is *neologism proof* if the equilibria giving rise to it are neologism proof. (From C1′ and C2′, an equilibrium is neologism proof precisely if all equilibria giving rise to the same outcome are neologism proof.)

REMARK 3.1: To see that the neologism-proof criterion does not always favor communication, consider another example adapted from Farrell [8], game Γ_2.

	action A	action B	action C
type a	1, 3	0, 0	2, 2
type b	0, 0	1, 3	2, 2

Game Γ_2

The no-communication outcome is $o(\cdot) = C$, and a communication equilibrium outcome also exists, $o(a) = A$ and $o(b) = B$. The former is trivially neologism proof, as both Sender types get their maximum payoffs. The communication outcome is not neologism proof; relative to it, '$\{a, b\}$' is a credible neologism. The meaning of this neologism is, "I am not going to tell you my type," and if it is believed, the Receiver takes action C and both Sender types benefit relative to the equilibrium.

Our feeling is that a no-communication equilibrium probably would be played in Γ_2. The neologism-proof criterion properly rejects the communication outcome.

4. Problems with the Credible Neologism Concept

The credibility conditions C1′ and C2′ are sometimes too restrictive. For example, consider Γ_3.

	action A	action B	action C
type a	4, 3	3, 0	1, 2
type b	3, 0	4, 3	1, 2

Game Γ_3

This game has a no-communication outcome, $o(\cdot) = C$, and a communication outcome, $o(a) = A$ and $o(b) = B$. Both are neologism proof. In particular, a credible neologism does not exist relative to the no-communication outcome. The two neologisms that must be checked for credibility relative to it are, "I am type a" and "I am type b." If the Receiver hears and believes the former, he take action A. But type b as well as a would gain by this switch from C to A; hence, this neologism is not credible. Similarly, neither is "I am type b."

Yet, Γ_3 is no less a coordination game than is Γ_1; it seems clear that communication would occur. There is a strong intuitive argument that the Sender would make a speech that would be credible and serve to convert a no-communication equilibrium to a more desirable communication equilibrium. Consider two statements:

m_a = "I am type a, and if I were not I would say I was type b,"
m_b = "I am type b, and if I were not I would say I was type a."

In addition to making a claim as to the type, each of these statements also describes what claim the other type would make. The two statements are consistent with each other in so far as they describe the same talking strategy for the Sender (the truthful, separating one). Action A will result if m_a is uttered and believed, and action B will result if m_b is uttered and believed. Suppose it is understood that a no-communication equilibrium is to be played, so that everyone expects C to be the outcome. Then, if either statement would be believed if uttered, type a will prefer to utter m_a to either uttering m_b or to staying with the equilibrium. Type b will prefer to utter m_b to uttering m_a or to staying with the equilibrium. Thus, for example, when the Receiver hears m_a, he should not think that it might have been said by type b, since he should realize that type b would have done better by saying m_b. The two statements *taken together* are credible relative to a putative no-communication equilibrium.

The point is that inferences drawn from one statement should be consistent with the inferences that would have been drawn from alternative statements that could have been made. Rather than having a criterion for a possible statement to be credible, one should have a criterion for an entire set of possible statements to be credible.

The next example indicates another problem that arises when a credibility criterion applies to single statements. In the previous example, a statement lacked credibility until account was taken of another statement that could have been made. The opposite occurs in the following example: a statement appears credible until account is taken of other statements that could have been made. Game Γ_4 is adapted from Myerson [16] and Okuno-Fujiwara and Postlewaite [17].

	action A	action B	action C	action D
type a	4, 5	5, 4	0, 0	1, 4
type b	0, 0	4, 5	5, 4	1, 4
type c	5, 4	0, 0	4, 5	1, 4

Game Γ_4

The only equilibrium outcome of Γ_4 is the no-communication outcome, $o(\cdot) = D$. It is not neologism proof. Three neologisms are credible relative to it:

m_{ab} = "my type is either a or b,"
m_{bc} = "my type is either b or c," and
m_{ca} = "my type is either c or a."

Consider, for example, statement m_{ab}. If it were made and believed, the Receiver would take action B. This would give types a and b payoffs of 5 and 4, respectively, which exceed the payoff of 1 that they would get in an equilibrium. Type c would get only 0, making him worse off than in an equilibrium. This show that m_{ab} is a credible neologism.

We question a standard according to which these three statements are deemed credible. Suppose each is considered credible, and therefore would be believed if uttered. Then type a prefers most to utter m_{ab}, since it results in B, his most preferred action. Type b, however, prefers uttering m_{bc} (to get C). Type b should never make statement m_{ab}, since it is understood, according to the accepted credibility standard, that m_{bc} would be believed. The Receiver, knowing that his opponent is rational, should infer from m_{ab} that the Sender's type is a, not that it is equally likely to be either a or b, as the literal meaning of

the message requires. Similarly, the Receiver should infer from m_{bc} that the Sender is type b, and from m_{ca} that the Sender is type c. None of the three statements should be believed, contrary to the initial presumption that each is credible. This argument by contradiction indicates that not all of them should be considered credible.

A credibility standard should be a theory that passes the game theorist's usual test, namely, that it not cause itself to be contradicted when the players know that they are all acting according to it.

5. Announcement-Proof Equilibria

We now propose three related refinement criteria. The first two address the problems discussed in the preceeding section. The "strong announcement proof" criterion addresses the problem illustrated by Γ_3, and the "announcement proof" criterion addresses the problems illustrated by both Γ_3 and Γ_4. The third criterion, "weak announcement proofness," addresses both of these problems and, in addition, the so-called "Stiglitz critique," which is discussed below. Readers may disagree over which of these criteria is most sensible; our own weak preference is for the announcement-proof criterion.

As with the neologism-proof criterion, the starting point is a putative equilibrium to be tested, (τ,α,β), which is the received way of playing the game. It is this assumption that is to be tested by way of an indirect argument: assuming the putative equilibrium is expected to be played, is there an argument that the Sender can make, instead of playing his equilibrium strategy, that will convince the Receiver to depart from the equilibrium?[11]

The Sender makes the argument to deviate from the putative equilibrium by making a speech we shall refer to as an "announcement." Announcements are generalizations of Farrell's neologisms. An announcement is believed if it satisfies a credibility criterion. Our desideratum for an announcement to be credible is that the belief it asks the Receiver to adopt should be rational in a sense typically adopted in game theory. The Receiver should

---

[11] Thus, it is not common knowledge that the putative equilibrium will be played, for then it *would* be played. See, e.g., Binmore [4] and Bicchieri [3] for discussions on the use of counterfactual putative equilibria in refinement theories.

realize, after he hears the announcement, two things: (1) that the Sender had known when he made the announcement that *any* of the announcements satisfying the credibility criterion would have been believed, and (2) that the Sender knew when he made the announcement that he could have obtained his equilibrium payoff by staying with the equilibrium. Thus, in order for an announcement to be credible, its literal meaning should convey the information that the types claiming to make the announcement prefer it to the equilibrium and, in addition, to any other announcement satisfying the credibility criterion.

We now turn formal. An *announcement strategy* for the Sender in a game $\Gamma^1(\pi)$ is a pair $d = \langle \delta, D \rangle$. The set $D \subseteq T$ is a nonempty set of *deviant types*. The function $\delta: D \to \Delta M$ is a talking strategy for the deviant types. An *announcement* is a pair $\langle m, d \rangle$, where $d = \langle \delta, D \rangle$ is an announcement strategy and $m \in \delta(D)$, where $\delta(D)$ is the set of messages sent with positive probability according to δ by the types in D. The literal meaning of the Sender's announcement is,

> "My type t is in D, and I am sending message m according to strategy $\delta(\cdot \mid t)$. If I had been type s in D, I would have made an announcement that differed only in so far as m would have been chosen according to strategy $\delta(\cdot \mid s)$. If my type had not been in D, I would not have used announcement strategy d."[12]

The credibility of an announcement depends upon the putative equilibrium the players had expected would be played before they heard the announcement. For the following definitions, (τ, α, β) is the putative equilibrium, $d = \langle \delta, D \rangle$ is a given announcement strategy, and $\langle m, d \rangle$ is a given announcement.

This announcement is believed if the other players are convinced that every type $s \in D$ would have announced $\langle \hat{m}, d \rangle$, with $\hat{m}$ chosen according to $\delta(\cdot \mid s)$, and that no type not in D would have made such an announcement. In this case the Receiver updates his belief about the Sender's type to a belief $\hat{\beta}(m,d)$ defined, for all $t \in T$ and $m \in \delta(D)$, by

---

[12] If δ is a pure strategy, the announcement is essentially a partition $\{D_1, \ldots, D_k\}$ of the set D of deviant types, with the Sender stating which subset contains his type.

$$(5.1)\qquad \hat{\beta}(t \mid m,d) = \begin{cases} \dfrac{\pi(t)\delta(m \mid t)}{\sum_{s\in D} \pi(s)\delta(m \mid s)} & \text{if } t\in D \\ 0 & \text{if } t\notin D \end{cases}$$

Announcement $\langle m,d\rangle$ is *believed* if it causes the Receiver to adopt belief $\hat{\beta}(m,d)$.

If $\langle m,d\rangle$ is believed, the game played in the action stage is $\Gamma^0(\hat{\beta}(m,d))$, and the Receiver will play a strategy in its set of equilibria, $E^0(\hat{\beta}(m,d))$. Because this set may contain more than one equilibrium, it is not obvious how to specify what the Sender expects to happen when the announcement is believed. One approach is to assume that the Sender also suggests in his announcement an equilibrium in $E^0(\hat{\beta}(m,d))$ for the Receiver to play, which is an approach taken in Myerson [16]. We take a more conservative approach. According to it, the announcement is credible only if, when it is believed, even pessimistic deviant Sender types expect to gain by deviating, and even optimistic nondeviant types would expect to lose by deviating.

A *pessimistic* Sender type t who announces $\langle m,d\rangle$ when it is believable expects to receive

$$(5.2)\qquad \underline{u}_S(m,d \mid t) = \min\{u_S(\sigma \mid t) \mid \sigma\in E^0(\hat{\beta}(m,d))\}.$$

An *optimistic* Sender type t expects in the same circumstance to receive

$$(5.3)\qquad \bar{u}_S(m,d \mid t) = \max\{u_S(\sigma \mid t) \mid \sigma\in E^0(\hat{\beta}(m,d))\}.$$

The first credibility condition is that pessimistic deviant types should prefer the announcement to the equilibrium:

C1: $\underline{u}_S(m,d \mid t) \geq u_S(\tau,\alpha \mid t)$ for all $t\in D$ and $m\in\delta(\{t\})$,
with the inequality strict for some $t\in D$ and $m\in\delta(\{t\})$.

Next, optimistic nondeviant types should prefer the equilibrium to the announcement:

C2: $\bar{u}_S(m,d \mid t) \leq u_S(\tau,\alpha \mid t)$ for all $t\in T\setminus D$ and $m\in\delta(D)$.

The third condition is that the announcement strategy be internally consistent, i.e., that the messages designated by δ for each deviant type be optimal for that type given that any announcement of the form $\langle \hat{m},d \rangle$ would be believed. Our approach is again conservative:

C3: $\underline{u}_S(m,d \mid t) \geq \bar{u}_S(\hat{m},d \mid t)$ for all $t \in D$, $m \in \delta(\{t\})$, and $\hat{m} \in \delta(D) \setminus \{m\}$.[13]

DEFINITION 5.1: An announcement strategy $d = \langle \delta, D \rangle$, and the corresponding announcements $\langle m,d \rangle$, are *weakly credible relative to the equilibrium* (τ,α,β) if d satisfies C1-C3. If no announcement is weakly credible relative to (τ,α,β), then it and its outcome are *strongly announcement proof.*[14]

Conditions C1 and C2 are analogous to the credible neologism conditions C1′ and C2′ in Section 3. A credible neologism is essentially a weakly credible announcement in which all deviant types use the same talking strategy: if 'K' is a credible neologism, then $\langle \delta,K \rangle$ is a weakly credible announcement strategy if δ is a constant function.

Weakly credible announcements are not necessarily credible neologisms. Recall, for example, that no neologism is credible relative to a no-communication equilibrium of the game Γ_3: any claim the Sender might make to identify his type in this game would be a claim that both of his types would like to make and have believed, given that the no-communication outcome is the alternative. However, if we let $D = \{a,b\}$, m_a = "my type is a," m_b = "my type is b," $\delta(m_a \mid a) = 1$, and $\delta(m_b \mid b) = 1$, then $\langle m_a,d \rangle$ and $\langle m_b,d \rangle$ are weakly credible. So, the no-communication outcome in Γ_3 is not strongly announcement-proof, but, as is easy to see, the communication outcome is. Recall that one of our goals in developing a

---

[13] If a deviant type t has multiple optimal deviant messages, then C3 implies that his expected payoff is the same regardless of which strategy in $E^0(\hat{\beta}(m,d))$ is played after he makes an optimal announcement $\langle m,d \rangle$. That is, if $\delta(\{t\})$ contains more than one message, then whether t is a pessimist or an optimist he gets the same expected utility: $\underline{u}_1(d,m \mid t) = \bar{u}_1(d,\hat{m} \mid t)$ for any $m,\hat{m} \in \delta(\{t\})$, including $m=\hat{m}$. This insures that t is willing to randomize over the messages in $\delta(\{t\})$.

[14] Either all or none of the equilibria that give rise to a particular outcome are strongly announcement-proof, since conditions C1 – C3 refer to the equilibrium only through the interim utility levels it gives the Sender.

criterion different from the neologism-proof criterion was to obtain this distinction between equilibria in games like Γ_3.

Note that the definitions immediately imply that a strongly announcement-proof equilibrium is neologism proof.

We are not content with weak credibility. Recall Γ_4. There, a convincing theory predicting that the statements m_{ab}, m_{bc}, and m_{ac} are believable would contradict itself, as then each statement would be made by only one type of Sender, instead of by the two types who are supposed to make it.

More generally, suppose that $d = \langle \delta, D \rangle$ and $d' = \langle \delta', D' \rangle$ are both weakly credible relative to (τ,α,β). Suppose Sender type t is a member of both D and D', and that this type prefers the consequence of announcement strategy d' to that of d. Then, an accepted credibility criterion according to which both announcement strategies are believable would contradict itself: type t would not announce according to d because he prefers d', so the Receiver should place no probability on t if he hears an announcement of the form $\langle m,d \rangle$, contrary to $t \in D$. Therefore, we ask that an announcement d which satisfies C1-C3 should also satisfy the following condition:

C4: If $d' = \langle \delta', D' \rangle$ also satisfies C1-C3 relative to (τ,α,β), then
$\underline{u}_S(m,d \mid t) \geq \bar{u}_S(m',d' \mid t)$ for all $t \in D \cap D'$, $m \in \delta(\{t\})$, and $m' \in \delta'(\{t\})$.

DEFINITION 5.2: An announcement strategy $d = \langle \delta, D \rangle$, and the corresponding announcements $\langle m,d \rangle$, are *credible relative to the equilibrium* (τ,α,β) if d satisfies C1-C4. An equilibrium relative to which no announcement is credible is *announcement proof*, as is its outcome.

It is immediate that a strongly announcement-proof equilibrium is announcement proof.

The force of condition C4 can be seen in example Γ_4. Each of the credible neologisms m_{ab}, m_{bc}, and m_{ac} corresponds to a weakly credible announcement strategy relative to the equilibrium outcome. But each one is kept by another from satisfying C4, so that none of them is credible. The outcome is announcement proof but not neologism proof.

Recalling that in Γ_3, the no-communication outcome is neologism proof but not announcement proof, we see that neither criterion implies the other.

Condition C4 can lead to the following possibility: an announcement strategy can be declared not credible because of the existence of another announcement strategy which is itself not credible. Cycling can even occur, as in Γ_4. We are not overly concerned; so as not to reject too many equilibria, we prefer a strict to a weak standard of credibility.

Even so, the announcement-proof criterion, like the neologism-proof criterion, can reject all equilibria. The following game, Γ_5, illustrates this and leads us into a discussion of the "Stiglitz critique."

	action *A*	action *B*	action *C*
type *a*	3, 3	1, 0	2, 2
type *b*	1, 0	0, 3	2, 2

Game Γ_5

The only equilibrium outcome of Γ_5 is the no-communication outcome, $o(\cdot) = C$. The neologism '{*a*}' is credible relative to this outcome: the Receiver takes action *A* if he believes '{*a*}', which would benefit only type *a* relative to *C*. So this game has no neologism-proof equilibria. Since '{*a*}' is a credible neologism, announcement strategies in which the deviating set of types is D = {*a*} are weakly credible, and they give *a* his highest payoff, 3. And only they are weakly credible: type *b* could only lose by a change from *C*, so that in no weakly credible announcement strategy is *b* a deviating type. Consequently, announcement strategies in which the deviating set of types is D = {*a*} satisfy C4 in addition to C1-C3, and so are credible. This shows that the equilibrium outcome is not announcement proof.

The lack of a general existence theorem for announcement-proof equilibria does not overly concern us, since our intent is not to provide a theory of play which is applicable to all games in all contexts. A fallback prediction for Γ_5 is simply that some equilibrium will be played, so that *C* will be observed. However, it seems likely to us that type *a* will convince the Receiver to take action *A*, resulting in nonequilibrium play.

That said, we now describe an argument, credited to Stiglitz by Cho and Kreps [5], to the effect that type a should not be able to make a convincing announcement in Γ_5. We present it in a way that highlights the assumptions we think it requires.

Consider an equilibrium of Γ_5 in which both types of Sender send message m. The argument against type a being able to make a convincing announcement that identifies himself and upsets the equilibrium is by way of contradiction. The assumption to be contradicted has five parts and concerns the announcement, "I am type a":

(i) the Receiver would believe the announcement if it were made;
(ii) the Sender knows (i);
(iii) the Receiver knows (ii);
(iv) the Sender knows (iii); and
(v) the Receiver knows (iv).

Given (ii), type a should make the announcement to get his best action, A. Because of (iii), the Receiver can deduce that a will make the announcement. Hence, if he receives the equilibrium message m instead of the announcement, he should infer that the Sender's type is b and therefore take action B. But then, because of (iv), Sender type b can deduce that m results in B and the announcement results in A. Type b should therefore also make the announcement. Then, because (v) implies that the Receiver can deduce all this, he should deduce that type b as well as a makes the announcement. He should therefore not believe the announcement, contrary to (i).

This critique leads to the conclusion that an announcement may not be credible even if it satisfies C1-C4. However, given the refinement rationale described at the beginning of the Section, the critique is not convincing. Recall that the basis of this rationale is a putative equilibrium that is expected to be played. So, in Γ_5, how could the Receiver be expecting an equilibrium to be played and, at the same time, be realizing that he would believe the announcement, "I am type a," and that both Sender types know that he would believe it? Assuming the Receiver has this realization before the game begins contradicts the assumption that he is expecting the equilibrium to be played, so that the hypothesis

that the equilibrium is putative should be rejected at this point.[15] The refinement rationale requires that the Receiver realize that an announcement is believable *only* after it has been made, contrary to (iii)-(v).

The Stiglitz critique is more compelling if one adopts the view that all players know beforehand that announcements are possible, and the nature of the believable announcements. In this case, the critique suggests that conditions C1-C4 are too weak to satisfactorily define credibility. The difficulty stems from the fact that if a subset D of types can successfully identify itself, then it may necessarily be in the interests of some of the remaining types to be thought to be in D, rather than be known to be in $T\backslash D$. However, if an alternative equilibrium exists in which the types in D send diffent messages than do types not in D, then the Sender can rationalize an announcment by D as being an announcement that this alternative equilibrium is being played, as well as that the Sender's type is in D. Given this hypothesis, the types not in D would not want to make the announcement.[16]

This discussion can be formalized into a weak announcement-proof criterion which is more immune to the Stiglitz critique. Recall that if an announcement strategy $d = \langle \delta, D \rangle$ satisfies C1–C3 relative to a putative equilibrium (τ,α,β), then the speech of the Sender which is associated with an announcement $\langle m,d \rangle$ informs the Receiver that the Sender's type is in D, and that he has sent m according to δ. If d also satisfies the following condition, then the speech can be considered to also include a declaration that the Sender has chosen to play an equilibrium $(\hat{\tau},\hat{\alpha},\hat{\beta})$ instead of (τ,α,β):[17]

---

[15] This response to Stiglitz's critique is similar to that of Cho and Kreps [5].

[16] As a referee observes, the Receiver may still not be confident that the Sender's type is in D when he hears the announcement, since there may be yet another equilibrium in which types in $T\backslash D$ send the same messages as do types in D. (Actually, this observation applies to any equilibrium in which some types separate, since there is always another equilibrium in which they pool.)

[17] These considerations regarding C3A are similar to those which motivate the "undefeated" criterion proposed in Okuno-Fujiwara and Postlewaite [17].

C3A: there exists a talking strategy $\hat{\delta}: T\backslash D \to \Delta M$ such that $\delta(D) \cap \hat{\delta}(T\backslash D) = \varnothing$, an action strategy $\hat{\alpha}$, and a belief function $\hat{\beta}$ such that $(\hat{\tau},\hat{\alpha},\hat{\beta})$ is an equilibrium, where $\hat{\tau} = (\delta,\hat{\delta})$.

To handle the cycling problem, condition C4 should apply to announcement strategies satisfying not only C1-C3, but also C3A. Specifically, C4 should be revised to the following:

C4': If $d' = \langle \delta', D' \rangle$ also satisfies C1-C3 and C3A relative to (τ,α,β), then $\underline{u}_S(m,d \mid t) \geq \bar{u}_S(m',d' \mid t)$ for all $t \in D \cap D'$, $m \in \delta(\{t\})$, and $m' \in \delta'(\{t\})$.

DEFINITION 5.3: An announcement strategy $d = \langle \delta, D \rangle$, and the corresponding announcements $\langle m,d \rangle$, are *strongly credible relative to the equilibrium* (τ,α,β) if d satisfies C1-C3, C3A, and C4'. An equilibrium relative to which no announcement is strongly credible is *weakly announcement-proof*, as is its outcome.

Note that the definitions imply that all announcement-proof equilibria are also weakly announcement-proof.

If a game has a unique equilibrium outcome, as does Γ_5, then no Sender type prefers one equilibrium over another, so that every equilibrium is weakly announcement-proof. Still, some games have no weak announcement-proof equilibria, as the following game illustrates:

	action *A*	action *B*	action *C*	action *D*
type *a*	6, 6	0, 0	0, 0	5, 5
type *b*	0, 0	6, 6	0, 0	5, 5
type *c*	0, 0	0, 0	−1, 6	5, 5

Game Γ_6

Observe that Γ_6 has two semi-pooling equilibrium outcomes. In the first one, type *a* obtains *A* and types *b* and *c* pool to obtain *D*. In the second one, type *b* obtains *B* and types *a* and *c* pool to obtain *D*. An equilibrium achieving the first semi-pooling outcome is an alternative equilibrium which type *a* can use to form a strongly credible announcement to upset any

putative equilibrium in which his payoff is less than 6. Similarly, an equilibrium achieving the second semi-pooling outcome can be used by type *b* to upset any putative equilibrium in which his payoff is less than 6. Thus, any weak announcement-proof equilibrium must give a payoff of 6 to both *a* and *b*. It is easy to show that Γ_6 has no such equilibrium, since such an equilibrium would have to be completely separating.

6. Coordination Results

One hope for a cheap-talk refinement is that it provide a criterion for when a no-communication equilibrium can be rejected in favor of a communication equilibrium which both players prefer. However, recall from Γ_3 that the neologism-proof criterion can accept a no-communication equilibrium which gives all types of Sender and the Receiver relatively low payoffs, even when a communication equilibrium exists which gives them all their highest possible payoffs. This cannot occur with the announcement-proof criterion, as the following proposition indicates.

PROPOSITION 6.1: *Suppose* (τ,α,β) *is an equilibrium giving each type of Sender his maximum possible payoff, assuming the Receiver chooses a Bayesian rational strategy:*

$$(6.1) \qquad u_S(\tau,\alpha \mid t) = \max\{u_S(\sigma \mid t) \mid \sigma \in E^0(\Delta T)\} \quad \textit{for each } t \in T.$$

Then (τ,α,β) *is strongly announcement-proof (and hence neologism proof, announcement-proof, and weakly announcement-proof). If, in addition, each Sender type is indifferent among the Receiver's optimal responses to any message he might send in the equilibrium* (τ,α,β)*, i.e., if*

$$(6.2) \qquad u_S(\tau,\alpha \mid t) = u_S(\sigma \mid t) \text{ for all } t \in T,\ m \in \tau(\{t\}), \text{ and } \sigma \in E^0(\beta(m)),$$ [18]

then each type of Sender is indifferent between (τ,α,β) *and any other announcement-proof equilibrium.*

---

[18] This holds if, e.g., the Receiver has a unique best reply to each equilibrium message.

PROOF: By (6.1), no announcement satisfies the strict preference in C1 relative to (τ,α,β); hence, (τ,α,β) is strongly announcement-proof.

Now assume (6.2). Suppose type t is not indifferent between (τ,α,β) and another equilibrium (τ',α',β'). Then, by (6.1), t strictly prefers, and all other types weakly prefer, (τ,α,β) to (τ',α',β'). Consider the announcement strategy $d = \langle \tau, T \rangle$. If it is believed, then the least it can give each type is, by (6.2), $u_S(\tau,\alpha \mid t)$; hence, d satisfies C1 relative to (τ',α',β'). C2 is satisfied vacuously — there are no nondeviating types. C3 is satisfied because of (6.2) and the fact that (τ,α,β) is an equilibrium. As (6.1) implies that no type could prefer another announcement strategy, d satisfies C4. So d is credible relative to (τ',α',β'), which is therefore not announcement-proof. ////

Proposition 6.1 shows that the announcement-proof criterion selects, if it exists, an equilibrium which has the property that it gives each type of Sender his maximum possible payoff. This property is very strong. The remaining propositions concern equilibria which satisfy the weaker property of giving each type of Sender his maximum equilibrium payoff; such equilibria are weakly preferred by each type of Sender to every other equilibrium. We refer to such equilibria as *Sender (interim Pareto) dominant.* (Of course, a Sender dominant equilibrium exists in only some games.)

It is easy to see that a Sender-dominant equilibrium need not be announcement-proof. Consider a game like Γ_5, which has no announcement-proof equilibrium, but has a unique equilibrium outcome. The uniqueness of the equilibrium outcome implies, trivially, that all equilibria are Sender-dominant, but none are announcement-proof.

The reason a Sender-dominant equilibrium need not be announcement-proof is that the existence of a credible announcement strategy does not imply the existence of an equilibrium preferred by all the types of Sender. Only if one could delete the non-deviant types from the game would the announcement strategy constitute an equilibrium that "all" types of Sender preferred to the putative equilibrium. Thus, the payoffs the deviant types can get by making a credible announcement can exceed any of their equilbrium payoffs. With respect

to the announcement-proof criterion (and the strong announcement-proof criterion), Payoffs which are not achieved in any equilibrium are relevant.

On the other hand, only equilibrium payoffs are relevant for the weak announcement-proof criterion. Because of condition C3A, a strongly credible announcement exists relative to a putative equilibrium only if another equilibrium exists which the deviant types prefer to the putative equilibrium. Thus, if the putative equilibrium is Sender-dominant, no strongly credible announcement exists and the equilibrium is weakly announcement-proof. This proves the following:

PROPOSITION 6.2: *An equilibrium is weakly announcement-proof if it is weakly preferred by each type of Sender to every other equilibrium.*

Our remaining proposition shows that if a Sender-dominant equilibrium is also separating, then it satisfies the strongest of the criteria we have considered, strong announcement-proofness.

DEFINITION 6.1: An equilibrium (τ,α,β) is *separating* if no two types ever send the same message: for all $s, t \in T$, $\tau(\{t\}) \cap \tau(\{s\}) = \varnothing$.

Note that the Receiver prefers a separating equilibrium to all other equilibria, since "more information is better." Thus, a separating Sender-dominant equilibrium is fully interim Pareto dominant, i.e., the Receiver as well as each type of Sender weakly prefers it to all other equilibria.

If a weakly credible announcement strategy exists relative to a separating putative equilibrium, then another equilibrium can be shown to exist which the deviant types prefer to the putative equilibrium. This new equilibrium is constructed by a straightforward splicing together of the deviant talking strategy, δ, into the putative equilibrium's talking strategy, τ. The non-deviant types send the same messages as they did before, in response to which the Receiver adopts the same beliefs as he did before. These beliefs are still rational because no message sent by a non-deviant type had been also sent by a deviant type — this is the crucial property that separation implies. Thus, since no equilibrium is

preferred to a Sender-dominant equilibrium by any of the types, no weakly credible announcement strategy exists relative to a separating Sender-dominant equilibrium. This is the gist of the proof of the following proposition.

PROPOSITION 6.3: *A separating equilibrium weakly preferred by each type of Sender to every other equilibrium is strongly announcement-proof (and hence neologism proof, announcement-proof, and weakly announcement-proof).*

PROOF: Assume (τ,α,β) satisfies the hypothesis but is not strongly announcement-proof. Then an announcement strategy $d = \langle \delta, D \rangle$ exists which is weakly credible relative to (τ,α,β). Because (τ,α,β) is separating, the set of messages used with positive probability, according to τ, by the types in D is disjoint from the set of messages used with positive probability by the types not in D:

$$\tau(D) \cap \tau(T \backslash D) = \varnothing. \tag{6.3}$$

Also, we can assume that no message is used by both a deviating type in the announcement talking strategy δ, and by a non-deviating type in the equilibrium talking strategy τ:

$$\delta(D) \cap \tau(T \backslash D) = \varnothing.^{19} \tag{6.4}$$

Now, define a new talking strategy τ', and a new belief function β', by,

$$\tau'(m \mid t) = \begin{cases} \delta(m \mid t) & \text{if } t \in D \\ \tau(m \mid t) & \text{if } t \in T/D, \end{cases} \tag{6.5}$$

$$\beta'(t \mid m) = \begin{cases} \hat{\beta}(t \mid m,d) & \text{if } m \in \delta(D) \\ \beta(t \mid m) & \text{if } m \notin \delta(D). \end{cases} \tag{6.6}$$

---

[19] Because M is countably infinite, if (6.4) does not hold, we can choose another equilibrium giving rise to the same outcome, and another weakly credible announcement strategy, for which it does hold. (For example, one can always choose an equilibrium τ that uses only odd-numbered messages, and a δ that uses only even-numbered messages.) This is legitimate, since an equilibrium is strongly announcement-proof if and only if all the equilibria that give rise to the same outcome are strongly announcement-proof (recall footnote 14).

Then, (6.3) and (6.4), together with the fact that the original belief function β is a Bayesian update of the prior using τ, imply that β' is a Bayesian update of the prior using τ'. Let σ be a selection from the correspondence E^0, so that $\sigma(\pi') \in \Delta A$ is an optimal mixed action for the Receiver when he believes the Sender's type has the distribution $\pi' \in \Delta T$. For each $m \in M$, define

$$(6.7) \qquad \alpha'(m) = \begin{cases} \sigma(\beta'(m)) & \text{if } m \in \delta(D) \\ \alpha(m) & \text{if } m \notin \delta(D). \end{cases}$$

Then, because (τ,α,β) is an equilibrium, C1-C3 imply that (τ',α',β') is also an equilibrium. But C1 also implies that some type $t \in D$ strictly prefers (τ',α',β') to (τ,α,β), contrary to the hypothesis. ////

We conclude with an example showing that even if one equilibrium interim Pareto dominates all other equilibria, one of the latter can still be announcement-proof. Consider game Γ_7.

	action A	action B	action C	action D	action E	action F	action G
type a	2, 5	5, 4	0, 0	1, 4	4, 6	3, 0	3, 0
type b	0, 0	2, 5	5, 4	1, 4	3, 0	4, 6	3, 0
type c	5, 4	0, 0	2, 5	1, 4	3, 0	3, 0	4, 6

Game Γ_7

The no-communication outcome is $o(\cdot) = D$, and the only other equilibrium outcome is separating: $o(a) = E$, $o(b) = F$, and $o(C) = G$. The separating outcome is preferred by all Sender types and, therefore, is announcement-proof. Yet, the no-communication outcome is also announcement-proof. Just as in Γ_4, there are too many conflicting weakly credible announcement strategies. The neologisms m_{ab}, m_{bc}, and m_{ac} in which pairs of types pool (see the discussion of Γ_4) are weakly credible in Γ_7 relative to outcome $o(\cdot) = D$. So is a completely separating announcement strategy $d = \langle \tau, T \rangle$ in which the Sender reveals his type and thereby obtains the outcome $o(a) = E$, $o(b) = F$, and $o(C) = G$. But type a prefers m_{ab} to d

and to m_{ac}, type b prefers m_{bc} to m_{ab}, and type c prefers m_{ac} to m_{bc}. Each kind of weakly credible announcement strategy is not credible because one of its deviating types prefers another weakly credible announcement strategy in which it also deviates. So both outcomes are announcement-proof. (Only the communication outcome is strongly announcement-proof.)

Postscript: Credible Message Equilibria

We recently became aware of Rabin [19], which proposes a "credible message" refinement criterion for Sender-Receiver cheap-talk games. We briefly comment on this concept and its relationship to announcement-proofness here.

The distinguishing and nicest feature of Rabin's criterion is that it makes no reference to a putative equilibrium. His analysis is therefore immune to criticisms regarding counterfactuals (see, e.g., Binmore [4]). Furthermore, an analysis of communication which makes no reference to a putative equilibrium can, in principle, be applied jointly with a variety of solution concepts. Rabin, in fact, applies his concept to rationalizable strategies as well as to equilibria.

By specifying which messages are credible without reference to an equilibrium, Rabin avoids the nonexistence problem to which anouncement-proof (and neologism-proof) equilibria are subject. Existence comes from the fact that messages which we would view as credible in the presence of a putative equilibrium, fail Rabin's credibility test and, indeed, may not seem particularly credible in the absence of such reference. If one takes the position that a putative equilibrium is relevant, then Rabin's criterion can admit too many (unintuitive) equilibria. (Rabin agrees that his criterion may not be selective enough; see the example and discussion on page 158 of Rabin, [19].)

Nonetheless, announcement-proof equilibria need not satisfy Rabin's criterion. It is easy to show that neither criterion implies the other.

Appendix

Game Γ_1

	action A	action B	action C
type a	2, 3	0, 0	1, 2
type b	0, 0	2, 3	1, 2

Equilibrium Outcomes	Credible Neologisms	Credible Announcements
$o(\cdot) = C$	'$\{a\}$'	$D = \{a\}$, $\delta(a) = 'a'$
	'$\{b\}$'	$D = \{b\}$, $\delta(b) = 'b'$
		$D = \{a, b\}$, $\delta(a) = 'a'$, $\delta(b) = 'b'$
$o(a) = A$, $o(b) = B$	None	None

Game Γ_2

	action A	action B	action C
type a	1, 3	0, 0	2, 2
type b	0, 0	1, 3	2, 2

Equilibrium Outcomes	Credible Neologisms	Credible Announcements
$o(\cdot) = C$	None	None
$o(a) = A$, $o(b) = B$	'$\{a, b\}$'	$D = \{a, b\}$, $\delta(a) = \delta(b)$

Game Γ_3

	action A	action B	action C
type a	4, 3	3, 0	1, 2
type b	3, 0	4, 3	1, 2

Equilibrium Outcomes	Credible Neologisms	Credible Announcements
$o(\cdot) = C$	None	$D = \{a, b\}$, $\delta(a) = 'a'$, $\delta(b) = 'b'$
$o(a) = A$, $o(b) = B$	None	None

Game Γ_4

	action A	action B	action C	action D
type a	4, 5	5, 4	0, 0	1, 4
type b	0, 0	4, 5	5, 4	1, 4
type c	5, 4	0, 0	4, 5	1, 4

Equilibrium Outcomes	Credible Neologisms	Credible Announcements
$o(\cdot) = D$	'$\{a, b\}$' '$\{a, c\}$' '$\{b, c\}$'	None

Game Γ_5

	action A	action B	action C
type a	3, 3	1, 0	2, 2
type b	1, 0	0, 3	2, 2

Equilibrium Outcomes	Credible Neologisms	Credible Announcements
$o(\cdot) = C$	'$\{a\}$'	$D = \{a\}$, $\delta(a) =$ 'a'

Game Γ_6

	action A	action B	action C	action D
type a	6, 6	0, 0	0, 0	5, 5
type b	0, 0	6, 6	0, 0	5, 5
type c	0, 0	0, 0	–1, 6	5, 5

Equilibrium Outcomes	Credible Neologisms	Strongly Credible Announcements
$o(a) = A$, $o(b) = o(c) =$ D	'$\{b\}$'	$D = \{b\}$, $\delta(b) =$ 'b'
$o(b) = B$, $o(a) = o(c) =$ D	'$\{a\}$'	$D = \{a\}$, $\delta(a) =$ 'a'
$o(\cdot) =$ D	both of the above	both of the above

Game Γ_7

	action A	action B	action C	action D	action E	action F	action G
type a	2, 5	5, 4	0, 0	1, 4	4, 6	3, 0	3, 0
type b	0, 0	2, 5	5, 4	1, 4	3, 0	4, 6	3, 0
type c	5, 4	0, 0	2, 5	1, 4	3, 0	3, 0	4, 6

Equilibrium Outcomes	Credible Neologisms	Credible Announcements
$o(\cdot) = D$	'$\{a, b\}$' '$\{a, c\}$' '$\{b, c\}$'	None
$o(a) = E,\ o(b) = F,$ $o(c) = G$	None	None

REFERENCES

1. D. AUSTEN-SMITH, Interested experts and policy advice, Rochester, mimeo, 1990.
2. J. BANKS AND J. SOBEL, Equilibrium selection in signaling games, *Econometrica*, **55** (1987), 647-662.
3. C. BICCHIERI, Strategic behavior and counterfactuals, *Synthese*, **76** (1988), 135-169.
4. K. G. BINMORE, Modeling rational players I, *Econ. and Phil.*, **3** (1987), 179-214.
5. I.-K. CHO AND D. KREPS, Signalling games and stable equilibria," *Quart. J. Econ.*, **102** (1987), 179-221.
6. V. CRAWFORD AND J. SOBEL, Strategic information transmission, *Econometrica* **50** (1982), 1431-1452.
7. J. FARRELL, Credible neologisms in games of communication, MIT WP 386, 1985.
8. J. FARRELL, Meaning and credibility in cheap-talk games, Berkeley mimeo, forthcoming in *Mathematical Models in Economics*, ed. *M*. Demster, Oxford Univ. Press, 1990.
9. J. FARRELL AND R. GIBBONS, Cheap talk, neologisms, and bargaining, MIT WP 500, 1988.
10. F. FORGES, Equilibria with communication in a job market example, *Quart. J. Econ.*, **105** (1990), 375-398.
11. S. GROSSMAN AND M. PERRY, Perfect sequential equilibrium, *J. Econ. Theory* **39** (1986), 97-119.
12. E. KOHLBERG AND J. F. MERTENS, On the strategic stability of equilibria, *Econometrica*, **54** (1986), 1003-1038.
13. S. MATTHEWS, Veto threats: rhetoric in a bargaining game, Pennsylvania, CARESS WP #87-06, 1988.
14. S. MATTHEWS, Veto threats: rhetoric in a bargaining game, *Quart. J. Econ.*, **104** (1989), 347-369.
15. A. MCLENNAN, Justifiable beliefs in sequential equilibrium, *Econometrica*, **53** (1985), 889-904.
16. R. MYERSON, Credible negotiation statements and coherent plans, *J. Econ. Theory*, **48** (1989), 264-303.
17. M. OKUNO-FUJIWARA AND A. POSTLEWAITE, Forward induction and equilibrium refinement, Pennsylvania, mimeo, 1987.
18. T. R. PALFREY AND S. SRIVASTAVA, Efficient trading mechanisms with pre-play communication, Caltech WP 693, 1989.
19. M. RABIN, Communication between rational agents, *J. Econ. Theory*, **51** (1990), 144-170.
20. D. SEIDMANN, Effective cheap talk with conflicting interests, *J. Econ. Theory*, **50** (1990), 445-458.
21. J. C. STEIN, Cheap talk and the fed: a theory of imprecise policy announcements, *Amer. Econ. Rev.*, **79** (1990), 32-42.